Acerca de los corpora paralelos: el proyecto Intercorp

Petr Čermák
Universidad Carolina, Praga

Resumen. El artículo presenta el proyecto Intercorp, realizado en la Universidad Carolina de Praga, con el objetivo de crear un corpus paralelo de veinticinco idiomas del mundo. En él se exponen las razones de ser y los puntos de partida del proyecto. La base reunirá datos textuales en veinticuatro lenguas acompañados de sus equivalentes checos. El componente español contará como mínimo con 4.500.000 unidades léxicas. La concordancia resultante será accesible a través de Internet.

Palabras clave: corpus paralelo, proyecto Intercorp, lingüística contrastiva.

Abstract. The article introduces the Intercorp Project launched at Charles University, Prague with an aim to create a parallel corpus of 25 languages of the world. The resulting corpus, to be available on the internet, will contain textual data in 24 languages along with their Czech equivalents. The Spanish part will comprise at least 4,500,000 lexical units.

Keywords: parallel corpus, Intercorp Project, contrastive linguistics.

1. La necesidad de crear corpora paralelos

No cabe duda alguna de que los corpora lingüísticos, desde hace bastante tiempo, vienen ganando terreno en la lingüística como la fuente más fidedigna de material lingüístico. Además, se han desarrollado métodos para incluir información gramatical en el corpus, por lo que los corpora dejan de ofrecer una mera fuente de material desordenado, y su uso va extendiéndose a otros campos de la investigación lingüística.

Data de recepción: 31.07.2006. Data de aceptación: 20.10.2006.

Uno de estos campos lo constituye la lingüística contrastiva y el estudio de las diferencias entre idiomas. Hasta hace poco los investigadores tenían que recurrir a métodos poco exactos en su búsqueda de material lingüístico para estudiar las relaciones mutuas entre dos idiomas: la comparación de unos cuantos textos con sus traducciones, o su pura intuición lingüística. Este último método, obviamente, tiene escasa credibilidad científica, ya que la intuición lingüística lógicamente difiere incluso entre las personas muy cultivadas; por su parte, la comparación de textos se basaba en muestras no lo suficientemente exhaustivas y, en realidad, se limitaba a analizar unos cuantos idiolectos (es decir, tomaba en consideración tan solo la opinión de uno o más traductores). Los resultados de tales análisis difícilmente podían ser satisfactorios, de ahí que se estudien las posibilidades de aplicar la metodología de la creación de corpora de un único idioma al análisis de más idiomas.

2. Puntos de partida del proyecto INTERCORP

Este artículo pretende presentar el proyecto Intercorp, realizado en la Universidad Carolina de Praga, con el objetivo de crear un corpus paralelo de veinticinco idiomas del mundo. Existen varias razones para la puesta en marcha del proyecto:

1) La citada necesidad de buscar fuentes más seguras de material lingüístico para el estudio contrastivo de idiomas.

2) La Universidad Carolina de Praga, en concreto el departamento del Corpus de la Lengua Checa de su Facultad de Filosofía y Letras, dispone de un enorme corpus de la lengua checa. El corpus, llamado ČNK (véase su presentación en http://ucnk.ff.cuni.cz/), ofrece una base idónea para el futuro corpus paralelo, tanto por sus textos como por su metodología. Como se podrá ver más tarde la importancia del corpus checo es sustancial para el proyecto Intercorp.

3) Últimamente, varios estudios han puesto de manifiesto la insuficiencia de incluso los mejores diccionarios bilingües. Así, un estudio reciente de František Čermák, director del proyecto Intercorp (Čermák – Klégr, 2004), en el que analiza los equivalentes ingleses de las partículas checas *snad, možná, asi, nejspíše* (partículas que expresan posibilidad / aproximación, equivalentes a *posiblemente, a lo mejor,* etc.) y *by* (marca del condicional en checo), subraya la necesidad de utilizar los corpora (es decir, los corpora paralelos) como fuente de material lingüístico en lingüística contrastiva. Los autores, sirviéndose de un pequeño corpus paralelo checo-inglés, buscan equivalentes ingleses de las partículas mencionadas. Los resultados alcanzados se comparan con los equivalentes del diccionario bilingüe checo-inglés, de mayor extensión. La detallada comparación nos hace ver que, en algunos casos, sólo los 40 % de los equivalentes encontrados en el corpus aparecen en el diccionario. Los autores llegan a la conclusión de que este hecho no se debe a la calidad de los diccionarios (han trabajado con el mejor diccionario checo-inglés existente), sino a una realidad lingüística muy compleja que no se presta a la descripción fácil que lógicamente debe aparecer en un diccionario. Como conclusión, apuntan lo siguiente:

> "The results of the present contrastive study do point to the insufficiency of even the best dictionaries. (…) There is no doubt a relation between the fact that description of modality particles has been rather underdeveloped in Czech grammatical theory so far and the way they are treated in monolingual and, consequently, even bilingual dictionaries. As a result the dictionaries look upon them as isolated lexical items and the description of the range of their modal meanings is rather sketchy. This reductionist tendency is carried even further in the bilingual dictionaries where modal distinctions are generally swept under the carpet and presumably "universal" world-class corresponding equivalents are usually offered instead. However, it is evident that parallel corpora may do much better in offering information on usage, too, which is missing from dictionaries in most cases. Such specific and complex cases as modal and mood marking elements certainly require much more information. This should include the mapping of distribution (their place in the sentence), constraints, if any, etc. (…) It is evident that only a corpus with its profusion of contexts may contribute to the resolution of epistemic, deontic and alethic types of modality, the expression of which in language is extremely varied and complex. This applies even more when it comes to correspondence between two languages as both parts of the study have shown. In fact, both have indicated serious discrepancies between dictionary equivalents and the actual situation in texts in terms of the number, frequency and type of equivalents occurring in texts and their distribution (…). The results of the study go on to confirm the necessity of re-evaluation of the hitherto descriptions against the background of information provided by parallel corpora" (Čermák – Klégr 2004: 94-95).

Lo dicho sobre la modalidad vale, evidentemente, para la mayoría de los fenómenos lingüísticos. Los corpora paralelos no pretenden sustituir a los diccionarios bilingües, sino complementarlos: ofrecer una información más completa, incluyendo la influencia del contexto, etc.

3. Descripción general del proyecto

El proyecto INTERCORP tiene como objetivo establecer una base de datos compuesta por textos de 25 lenguas, entre ellas el checo, el español el francés, el italiano y el portugués. La base reunirá datos textuales en 24 lenguas acompañados de sus equivalentes checos. La concordancia resultante será accesible a través de Internet.

Es evidente que un proyecto tan amplio requiere la participación de especialistas de todas las lenguas que lo constituyen. Por eso participarán numerosos departamentos de la Facultad de Filosofía y Letras. El Instituto de Lenguas Románicas de la Facultad de Filosofía y Letras de la Universidad Carolina es el organismo encargado de la dirección y coordinación de los trabajos que tienen como fin la construcción de un conjunto de textos en francés, español, portugués e italiano, con sus correspondientes en checo.

El corpus así creado será utilizado para fines exclusivamente científicos (investigación en materia de lingüística y teoría de la literatura), así como para trabajos de lexicografía y

para las necesidades relacionadas con la enseñanza de lenguas extranjeras en el ámbito universitario. Por otra parte, los corpora permiten poner en evidencia los aspectos sintagmáticos de los fenómenos lingüísticos en toda su complejidad, lo que era, hasta hace poco, inimaginable.

En cuanto a metodología, y como ya se ha mencionado, los autores del proyecto se apoyan en la metodología desarrollada por el departamento del Corpus de la Lengua Checa de la Universidad Carolina (ČNK), sobre cuya base los investigadores han creado un corpus en checo (llamado, también, ČNK), así como los programas que permiten recuperar los datos que lo constituyen.

Las características fundamentales del proyecto son las siguientes:

— En una primera fase, se pretende recopilar un conjunto de material textual en versión electrónica, con una extensión que debería ser de 400.000 a 1.000.000 de palabras en función de la accesibilidad y del volumen de los datos. A continuación, seguirá su tratamiento informático, sin olvidar el desarrollo de los programas apropiados (en colaboración con especialistas en la materia). Este tratamiento informático de los datos abarca el alineamiento y una simplificación del etiquetado.

— A partir de esta primera fase, los corpora se irán aprovechando para una investigación continua. Los resultados de la investigación, así como de los resultados de talleres y conferencias, serán publicados en forma de libros, artículos y CD.

— En una segunda fase, se pretende crear un corpus suficientemente extenso, en función de las necesidades de las distintas secciones lingüísticas. Se calcula que será posible extender cada componente paralelo a un mínimo de 2.500.000 unidades léxicas. No obstante, queda claro que la situación en cada lengua es diferente (por ejemplo, en algunas lenguas existe un menor número de traducciones del/al checo, etc.).

— El objetivo final es la publicación de los corpora (autorizada por los prestatarios de los datos textuales) para fines de investigación llevada a cabo dentro del ámbito universitario; posteriormente, los autores del proyecto esperan hacer accesible el corpus en una página web, en función de sus posibilidades técnicas.

4. Descripción del componente español del proyecto

Desde hace un año, un grupo de hispanistas –profesores y estudiantes– de la Universidad Carolina viene recopilando textos en versión checa y española y adaptándolos a los requisitos del tratamiento informático. Su labor consiste ante todo en la búsqueda de textos idóneos, en el control detallado del grado de la correspondencia de las dos versiones y en su alineación. Las características sustanciales del futuro componente español del proyecto son las siguientes:

— En el proyecto Intercorp, el español se encuentra entre las lenguas "mayores". Este hecho no se debe solo a su prestigio natural, dado por su posición en el mundo y el número

de sus hablantes, sino también a la larga tradición de las relaciones culturales entre el mundo hispanohablante y la República Checa, que se ha manifestado, entre otras cosas, en una cantidad relativamente grande de traducciones de libros españoles e hispanoamericanos al checo. Todo parece indicar, pues, que la extensión mínima del componente paralelo, o sea 2.500.000 unidades léxicas, se sobrepasará sin problemas. Actualmente, el corpus provisional cuenta con unas 2.000.000 unidades léxicas.

— Los autores del proyecto pretenden que exista –si es posible– equilibrio entre las traducciones al checo y las traducciones a la segunda lengua. En el componente español, lo más probable es que no se logre este objetivo: la mencionada abundancia de las traducciones al checo contrasta con el número relativamente escaso de las traducciones al español.

— Un problema bastante grave lo constituye la calidad de la traducción de los textos, que difiere notablemente, especialmente en las traducciones del checo al español. Las traducciones se someten a un control previo y las traducciones inadecuadas no se incluyen en el proyecto. Dado que este problema atañe casi exclusivamente a las traducciones al español, aumenta, naturalmente, todavía más el desequilibrio entre las traducciones al checo y las traducciones al español.

— Otro de los objetivos de los autores del proyecto es –si es posible– un equilibrio temático del corpus: sería ideal que cada componente contara con un 50 % de los textos literarios (novelas, dramas, etc.) y con un 50 % de los demás textos (documentos oficiales, instrucciones, textos especializados, etc.). No obstante, el español no ofrece suficientes traducciones del segundo tipo de textos; lo más probable es que prevalezcan textos literarios en el componente español.

— Además de los criterios temáticos, el proyecto cuenta también con un criterio temporal. El corpus debería reflejar el estado actual de los idiomas, por eso solamente se incluyen textos de los siglos XX y XXI; en la mayoría de los casos, el criterio es todavía más estricto y se trabaja solo con los textos que proceden de los últimos 50 años.

— Los que preparamos el componente español del corpus contamos ante todo con los usuarios que se interesen por el estudio contrastivo de las lenguas checa y española. No obstante, cierta parte del corpus servirá también –de acuerdo con los objetivos del proyecto– para la comparación del español con otros idiomas que no sea checo, es decir, con un "tercer" idioma. El núcleo de esta parte lo constituirán textos de los autores checos más prestigiosos, cuyas obras han sido traducidas a varias lenguas, y también algunos documentos oficiales (por ejemplo, documentos oficiales de la Comunidad Europea). Además, en la última fase de la preparación del corpus pretendemos incluir también traducciones españolas de textos escritos en lenguas que no sea el checo (como es lógico la literatura inglesa y americana ofrece más traducciones a los demás idiomas; no obstante, hay otros autores, como Franz Kafka, que también han sido traducidos a numerosas lenguas). La utilidad del proyecto será, por tanto, mucho más grande, aunque solo en una de sus vertientes: los interesados podrán comparar, por ejemplo, construcciones inglesas con construcciones españolas.

5. Conclusión

Creemos que el proyecto que acabamos de presentar será de gran utilidad no solo para los hispanistas checos –aunque, lógicamente, serán estos los que más provecho sacarán de él– sino también para toda la comunidad lingüística internacional; además de suministrarles material lingüístico –ordenado y preparado para un análisis teórico– les dará cuenta de las posibilidades actuales de los corpora paralelos. Los interesados pueden dirigirse al autor del presente artículo (petr.cermak@ff.cuni.cz) o a la página web del proyecto (http://trnka.ff.cuni.cz/ucnk/intercorp/).[1]

Bibliografía

Blatná, R. – Petkevič, V. (eds.) (2005), Jazyky a jazykověda. Sborník k 65. narozeninám prof. Františka Čermáka, Praga, ÚČNK FF UK.

Čermák, F. – Klégr, A. (2004), "Modality in Czech and English. Possibility particles and the conditional mood in a parallel corpus", International Journal of Corpus Linguistics, Vol. 9:1. 83-95.

Čermák, F. – Klímová, J. – Petkevič, V. (eds.) (2000), Studie z korpusové lingvistiky, Praga, Karolinum.

Salkie, R. (2002), "Two types of translation equivalence" en Altenberg B. & Granger S. (eds.), Lexis in Contrast. Corpus-based approaches, Amsterdam/Philadelphia: John Benjamins, 51-71.

1 Esta nota forma parte de la investigación subvencionada por el Ministerio de Educación, Juventud y Educación Física de la República Checa (MSM0021620823).